METHOD AND SYSTEM FOR TRANSMITTING VOICE MESSGES

Technical Field

The present invention relates to a communications method and system which uses speech recognition technology to analyse a voice message so as to determine its intended destination.

Background to the Invention and Prior Art

Mobile packet based half-duplex voice messaging systems are known in the art. Referred to colloquially as "push-to-talk" (PTT) systems, they have been commercially available within the United States for some years, provided by Nextel Communications, under the service mark "Direct Connect".

Such PTT systems have also been developed to operate within an internet protocol (IP) environment, with voice over IP (VoIP) systems. In particular both General Packet Radio Services (GPRS) and Code Division Multiple Access (CDMA) based VoIP PTT systems are known in the art, such as those produced by Motorola (see http://www.motorola.com/mediacenter/news/detail/0,1958,3069_2512_23,00.html) and Qualccomm (see http://www.qualcomm.com/press/releases/2002/020111_qchat_voip.html).

When using a PTT system, usually a user will select the intended receiver from an address book list maintained on his own handset using a graphical interface and the device's own user controls, as is well known in the art. It is also known to provide for voice dialling of PTT services, however, and an example prior art device which provides such functionality is the pocket adapter produced by Cellport Systems Inc. of Boulder , CO, for the Motorola iDEN i1000 and i1000 plus mobile telephones. The user guide for the Cellport pocket adapter can be found at http://www.cellport.com/adapterguides/nextel_i1000_PAG.pdf. As set out therein, such voice dialling comprises the user speaking predetermined code words, followed by the identification (such as the number, but alternatively a speed dial code) of the receiver which the user wishes to connect to, before the voice message which the user wishes to send is spoken. For example, in the Cellport system, using voice dialling a user would speak the words "Cellport, dial, pound, pound, 6284". The adapter then repeats the recognised words "pound, pound, 6284", and then the connection process is performed. The user can then speak his message by pressing the PTT button in the usual way.

Even with such voice dialling functionality, however, there is still a separate "dialling phase", where the user must select the intended recipient, either by using a normal graphical interface, or by using the voice dialling interface, and it is not until such dialling phase has been completed and a connection established that the user may speak his first message. This separate dialling phase therefore introduces a delay in allowing a user to speak his message, and also necessitates additional user interaction with the device, either in the form of navigating the graphical displays, or by speaking in accordance with the voice dialling protocols.

Summary of the Invention

The invention aims to improve on the above described operation by removing the separate dialling phase from the user interface. More particularly, the invention makes use of speech recognition and associated technology to analyse a spoken message so as to identify an intended receiver for the message, and to transmit the message or at least a variant thereof (such as the text of the message as obtained by the speech recogniser) towards the intended recipient via a network. This allows a user to simply push the PTT button on his handset and immediately speak his message (preferably including within his message some indication of the intended recipient, such as a name or the like), without having to undergo a separate dialling phase beforehand.

In view of the above, from a first aspect there is provided a communications method comprising the steps of:-

receiving a voice message containing an utterance;

buffering the received message;

performing a speech recognition process on the received voice message to recognise the utterance contained therein;

determining, if possible, an intended receiver of the message in dependence on the recognised utterance; and

if an intended receiver was determined, transmitting the message to the determined intended receiver using a half-duplex communications service provided by a packet-switched network.

As set out above, the invention provides the advantage that no separate dialling phase need be undertaken before a user may speak the message. This enhances the communication experience for the user, and makes the half-duplex communications service more pleasant and user friendly to use.

In an embodiment of the invention, when the determining step determines one or more possible intended receivers from the recognised utterance, the method further preferably comprises the steps:

indicating the one or more possible intended receivers to a user; and

receiving a selection signal from the user indicating the one or more determined possible intended receivers to which the message should be transmitted.

Thus, where the determining step has identified one or more potential intended recipients for a message, clarification of which of the identified possible intended recipients may be sought from the user. Preferably, for ease of interface, the indicating step further comprises generating an audio speech prompt corresponding to the plurality of possible intended receivers; and outputting the generated audio speech prompt to the user. Such a further feature allows for an audio output prompt from the user device.

Moreover, in embodiments of the invention the speech recognition process is preferably performed only on a portion of the received voice message. Such a feature recognises that it is likely that the intended recipient of a message will be indicated probably at the start of a message (e.g. consider the message "Hi Roger, are you going to the pub this evening?" – the intended recipient ("Roger") is identified in the first phrase), and hence speech recogniser resources may be conserved by performing only that amount of recognition which is necessary.

Furthermore, in embodiments of the invention the further steps of: receiving an indication of the identity of a user who generated the message; and selecting a user-dependent speech grammar for use by the speech recognition process in dependence on the identity of the user are included. This allows a user-specific grammar to be used with the speech recognition process, which grammar may have encoded therein address book data and the like.

Additionally, embodiments of the invention may further comprise the steps of receiving a speech recognition activation signal from a user, wherein the speech recognition and determining steps are performed in dependence on the receipt of such a signal. Such functionality allows a user to explicitly indicate when a message is a message to a new recipient, and hence that the speech recognition and receiver determination steps should be performed. This further improves the efficiency of use of speech recogniser resources, and also improves the overall operation of the invention, as the speech recognition and receiver determination steps will only be performed on messages (typically first messages in a thread) for which there is a high likelihood that some sort of receiver identity such as a name or the like will be spoken therein, and

hence a correspondingly high likelihood that the intended recipient will be capable of determination.

Brief Description of the Drawings

Further features and advantages of the present invention will become apparent from the following description of an embodiment thereof, presented by way of example only, and by reference to the accompanying drawings, wherein:-

Figure 1 is an architectural system block diagram of an embodiment of the present invention;

Figure 2(a) is flow diagram illustrating the method steps involved in the embodiment of the invention; and

Figure 2(b) is a flow diagram continuing the flow diagram of Figure 2(a).

Description of the Embodiment

An embodiment of the present invention will now be described with respect to Figures 1, 2(a) and 2(b).

The embodiment of the invention provides a voice steered push to talk (PTT) service. More particularly, the embodiment of the invention is intended to provide a push to talk communication service which may use any of the PTT communications technologies already known in the art and discussed in the introductory portion to this specification, and which then adds thereto functionality which allows the push to talk communications to be directed to an intended recipient or destination without undergoing any explicit dialling phase. Instead, within embodiments of the invention, speech recognition is performed on the spoken messages and a speech grammar applied to determine intended recipients or destination of the message, the message then being forwarded to the intended recipient or destination thus determined.

In view of the above, Figure 1 illustrates an overall system architectural block diagram illustrating the main system elements of an embodiment according to the present invention. With reference to Figure 1, therefore, within the embodiment of the invention there is provided an audio router server 14 which is arranged to receive streamed digital audio signals carried by a PTT communication service on a network (not shown) from PTT-enabled handsets, as well as calling line identifier (CLI) information. The audio router server 14 is provided with an audio buffer 142, being a data storage medium such as RAM, a hard disk, an optical storage medium, or the like, and which is intended to store any received audio messages therein temporarily. Additionally provided by the

embodiment is a speech recognition server 18, which is arranged to receive digital audio from the audio buffer 142 at the audio router server together with the CLI information, and also to receive speech grammar and lexicon data for use in a speech recognition process, from an address book and grammar database 20. In use, the speech recognition server 18 runs a speech recognition application to apply a user specific grammar to the digital audio received from audio buffer 142, so as to recognise any spoken utterance therein, and determine an intended recipient. It should be noted that the speech recognition application run by the speech recognition server may be any speech recognition application presently known in the art, but preferably a speaker independent speech recognition application. Suitable speech recognition software which was available before the priority date and which may be used by the speech recognition server 18 in the present embodiment is Nuance 7, from Nuance Communications Inc, of 1005 Hamilton Court, Menlo Park, CA 94025.

The speech recognition server 18 is further arranged to pass a recognition result, being preferably a set of key-value pairs representing the values of particular grammar slots in the recognised speech together with their associated recognition confidence values, to a recipient determination server 16. The recipient determination server is arranged to receive the key-value pairs, and to take action appropriately dependent on the key value pairs returned by the recogniser, as will be described later. One of the possible actions which the recipient determination server can perform is to pass an address in the form of a Dialled Number Identifier (DNI) to the audio router server 14. Additionally, the recipient determination server 16 is further arranged to receive calling line identifier (CLI) data from user handsets (described later) and also to send shortlist information to user handsets, as will also be described later.

Returning to a consideration of the address book and grammar database 20, this stores, for each registered user of the system, a speech recognition grammar which encodes address book data relating to names of possible recipients and their respective DNIs. A separate user specific grammar is stored for each registered user. An example format for a grammar is shown below:-

```
Names
(
[
   ( bob ) { return("Bob Smith +447711123456") }
   ( peter jones ) { return("Peter Jones +447722123456") }
   ( pete ) { return("Pete Brown +447733123456") }
```

```
]
)


Phonemes:filler
[
ph1
ph2
...
ph41
]

Fillers:filler
[
  Phonemes
  @-@
]

EndCall
[
   end call
   over and out
]

Overall
[
  ( ?hi +Names:n ?(it's Bob) *Fillers) {<action "placecall">
<recipient $n>}
  *Fillers
  EndCall {<action "endcall">}
]
```

The creation of the dialling grammar may be by any of the well known techniques using either text (e.g. getting a copy of user telephone's address book) or combination of the address book for the number part and spoken input to define the name pronunciation.

The above description relates to the various servers which are preferably contained within or form part of a network providing the half duplex PTT communications service. In order to describe the operation of the embodiment in more detail, however, it is necessary also to describe the special features of the mobile user handsets adapted for

use with the present invention. Figure 1 also illustrates such a handset, in the form of mobile station A (10). Mobile station B (12) is also shown, but this may be a conventional PTT enabled handset, as is already known in the art.

Referring therefore to mobile station A (10), the mobile station A (10) is provided with a PTT audio client A (102), being the appropriate software to control mobile station 10 to provide conventional PTT functionality. That is, the PTT audio client A (102) enables the mobile station 10 to use a PTT service in a conventional manner. The main difference with the prior art, however, is that the PTT audio client A (102) is arranged to send the digitised audio data produced by the handset A to the audio router server 14, together with the calling line identifier of the mobile station 10.

Additionally provided within the mobile station 10 is a visual selector client A (104). This is a further software program which is arranged to interface with the recipient determination server 16 within the network, so as to send the calling line identifier (CLI) of the mobile station 10 thereto when a PTT call is first initiated, and also to receive a list of candidate recipient identities from the recipient determination server 16, in the event that more than one intended recipient is determined thereby. The visual selector client A (104) is further arranged to display such a short list of candidates to the user on a display of the mobile station A, and to permit the user to select the intended recipient. Selection information is then transmitted back to the recipient determination server 16.

Within the conventional mobile station B (12), a PTT audio client B (122) is provided, which is essentially the same as the PTT audio client A (102) provided in the mobile station 10. The PTT audio client B (122) is arranged to provide the mobile station B (12) with conventional PTT functionality, and the only difference between the PTT audio client B (122) and the prior art is that the PTT audio client B (122) is arranged to receive a PTT message from the audio buffer 142 which is part of the audio router server 14. In all other respects, the mobile station B12 may be conventional.

The operation of the embodiment of Figure 1 will now be described with respect to Figures 2(a) and (b).

In this example description of the operation of the embodiment, let us assume that mobile station A does not have any PTT calls in progress, and wishes to send a PTT message to mobile station B. That is, the present state of mobile station A is that it has not sent or received any PTT calls to any other station for at least a PTT timeout period (usually 20 seconds). In view of this, within the embodiment of the invention in order to initiate a call from mobile station A to mobile station B at step 2.2 user A presses the PTT button. The pressing of the PTT button on the mobile station A causes the PTT audio

client 102 to start running, as well as the visual selector client 104. At step 2.4 the visual selector client 104 connects to the recipient determination server 16, and sends the calling line identifier (CLI) of the mobile station A to the recipient determination server 16. Next, (or almost simultaneously or beforehand - the order of steps 2,4 and 2.6 is not important) the PTT audio client 102 connects to the audio router server 14, and starts streaming digitised audio to the audio router server 14. It is at this point, at step 2.6, that user A of mobile station A speaks the message which he wishes to be transmitted by the PTT service, and the mobile station A digitises and packetises the message for streaming in the audio stream to the audio router server 14. Such digitisation and packetisation is well known in the art. In addition to streaming the audio to the audio server 14, the PTT audio client 102 also sends the calling line identifier (CLI) of the mobile station A to the audio router server 14.

At step 2.8, the audio router server buffers the received audio stream in the audio buffer 142, and also forwards a copy of the audio stream to the speech recognition server 18, at step 2.10. At the same time, the audio router server 14 also sends the mobile station A calling line identifier to the speech recognition server 18.

Next, at step 2.12, the speech recognition server 18 uses the received calling line identifier of the mobile station A to access the address book and grammar database 20, so as to retrieve therefrom the specific user grammar which is stored therein for the mobile station A. It will be appreciated that the speech recognition grammar and lexicon is stored in the address book and grammar database 20 indexed by CLI, to allow for the grammar and lexicon specific to the mobile station A to be retrieved.

Next, at step 2.16 the speech recognition server 18 performs a speech recognition process on the audio stream received from the audio router server 14. Note that the speech recognition server 18 may perform speech recognition on the received audio stream as the stream is received, or alternatively may wait until silence is detected in the stream before commencing recognition [or the end of the stream when the PTT button is released]. This choice will depend on the precise speech recognition software chosen for use within the speech recognition server 18. The speech recognition process performed by the speech recognition server 18 acts to recognise the user utterance contained within the audio stream received from the audio router server 14, using the recognition grammar for the user to guide the recognition process. Within the embodiment the speech recognition server then returns key information to the recipient determination server via Nuance NL slots (when the Nuance 7 recogniser mentioned earlier is used), so, for example, for the utterance "over and out" the recogniser would return a key value pair

of *action=endofcall*, while for "Hi Bob, it's Bob" the recogniser would return two key value pairs: *action=placecall* and *recipient="Bob Smith +447711123456"*, as determined by the user grammar. With the key-value pairs the recogniser also provides a recognition confidence value indicative of how confident it is of any particular recognition leading to a particular key-value pair being correct. Such recognition confidence values are well known in the art.

Having performed the recognition, and output the key-value pairs and confidence values to the recipient determination server, at step 2.20 an evaluation is performed by the recipient determination server on the confidence values of the returned key-value pairs. Here, the confidence values may be compared with a threshold value or the like, and if the evaluation indicates that the recogniser is confident of the results, then processing may proceed to step 2.22. Here, at step 2.22 the recipient determination server 16 sends the DNI(s) of the determined intended recipient(s) and obtained from the received key-value pairs to the audio router server 14, and also, as confirmation, to the visual selector client 104 in the mobile station A. Then, at step 2.24 the audio router server transmits the buffered audio message from the audio buffer 142 to the receiver(s) identified by the DNI(s) received from the recipient determination server, using the PTT communications service. In this case, let us assume that the DNI(s) received from the recipient determination server identifies mobile station B, in which case the audio router server streams the audio message from the audio buffer 142 to the PTT audio client 122 in the mobile station B, over the usual PTT enabled network. At the same time, at step 2.26 the visual selector client 104 at the mobile station A displays the determined DNI(s) to the user A on the display of the mobile station A, as confirmation that the message has been forwarded properly. At that point, therefore, a PTT call has been set up by the audio router server between the mobile station A and the mobile station B, and PTT communications may then continue in a conventional manner.

It is important to note here that it is possible for the speech recognition server to confidently recognise two or more intended recipients, when two or more recipient identifier's are spoken (consider here the message "Pete, Bob, its Dave here", in which case both Pete and Bob are intended recipients). Due to the recognition grammar both or all of the intended recipient's DNIs may be returned, and due to the confident recognition of both or all it becomes clear that the message was intended for both or all recipients. In such a case the recipient determination server controls the audio router server to set up a group PTT call, to each of the determined intended recipients (Pete and Bob in the example). This feature therefore allows for calling groups for group calls to be defined

dynamically, by simply indicating in the message the names of each of the intended recipients which are to be parties to the group call.

Returning to step 2.20, if the evaluation performed thereat does not indicate that there is a confident result, then the recipient determination server performs a further evaluation at step 2.28, to determine whether or not there are one or more non-confident results returned from the speech recogniser. If it is the case that no key-value pairs were returned, then the recognition process has failed. In this case the recipient determination server sends a message at step 2.36 to the visual selector client 104 at the mobile station A that recognition has failed, and a recognition failed message is then displayed to the user at the mobile station A. In such a case, the user A must then select the intended recipient for his message using the conventional graphical user interface.

On the contrary, however, if the evaluation of step 2.28 indicates that there are one or more non-confident results, then the user is invited to confirm the one or more non-confident results. Therefore, at step 2.30 the recipient determination server 16 sends a list of the one or more non-confident results to the visual selector client 104 at the mobile station A. Visual selector client 104 then displays the received list to the user on the display of the user on the display of the mobile station A, and at step 2.32 the user selects the intended recipient from the list. The visual selector client 104 then sends the selection information back to the recipient server 16 at step 2.34. The recipient determination server receives the selection information and then returns to step 2.22 wherein the selected DNI (or DNIs where multiple intended recipients are selected) is sent to the audio router server. Processing then proceeds to step 2.24 and onwards as described previously.

In view of the above description, therefore, it will be seen that the embodiment of the invention allows for convenient PTT communications without the user having to undergo a specific dialling phase, and waiting for a subsequent connection. In this respect, the invention makes use of the inherent latency in PTT communications, and in particular VoIP implementations thereof, and exploits that latency to perform speech recognition and subsequent intended recipient determination on the message, to allow for automatic recipient or destination selection. The invention therefore provides an enhanced user experience over and above that provided by the conventional PTT communications services known in the art.

It will be appreciated that various modifications may be made to the described embodiment to produce further embodiments. For example, in a further embodiment in order to initiate the recognition and recipient determination process, the user A at the mobile station A must send an activation signal from the mobile station A, for example by

double clicking the PTT button. Such a "double click" would cause the visual selector client A (104) to send an activation signal to the recipient determination server 16, which in turn sends an activation signal to the speech recognition server 18. Such an explicit activation operation may be beneficial to prevent the invention operating in unwanted circumstances.

As another variant, in further embodiments the user A may make a group call using the invention. Group calling using PTT is known *per se* in the art, and is included for use within embodiments of the invention by including within each user address book a group name, together with the associated telephone numbers which form part of the group. In operation, the user A speaks the group name, which is then recognised by the speech recognition server, and the stored group name applied to the user grammar to determine the DNIs for the group. If the group name is recognised, then the recipient determination server sends each of the DNIs belonging to the group to the audio router server 14, which then connects the group PTT call in a conventional manner.

In a further embodiment, the speech recognition server is arranged to recognise only the first few seconds of a message, so as to conserve speech recogniser resources. This feature is based on the premise that for most greetings the recipient name will be said within such a limit (consider the greetings: "Hello, Pete"; "Hi Bob,"; "Good Morning, Pete" etc.). Recogniser time limits of between 3 and 5 seconds should suffice for this purpose.

Regarding the selection of intended recipients in the event of non-confident results, in the embodiment above we describe the visual selector client displaying the shortlist to the user for visual selection. In other embodiments, however, the recipient determination server may include a speech synthesiser program which is used to generate audio prompts relating to the selections available, which are then routed to the PTT audio client 102 at the handset for playing to the user. Note that this may be performed simultaneously with the display of the shortlist by the visual selector client, such that the selections are presented by both audio and visual interfaces, or alternatively may replace the visual selection. In order to allow for spoken selection by the user of an intended recipient (for example, the user speaks "Yes" when the intended recipient is read out, and/or (optionally) "No" when the name of a non-intended recipient is played, or alternatively the user speaks "Bob Smith" to distinguish between Bob Smith and Bob Jones, previously referred to simply as "Bob"), the PTT audio client may transmit any user response to the speech recognition server via the audio router server for recognition of the responses.

Finally, as a further optional feature the audio router server can be arranged in other embodiments of the invention to trim the audio which has been recognised and used to select the intended recipient from the message, and to transmit only that part of the message which was not used for the intended recipient determination.

Whilst the above description concentrates on the operation of the invention prior to the setting up of a PTT call, in another mode the embodiments of the invention can operate during a PTT call (that is, -within the PTT timeout period when messages are expected to be travelling back and forth between two or more parties to a call) to detect a predetermined "end-call" phrase, such as "Over and out", or "End Call", and to operate to close down the call. In this mode of operation the audio stream routed through the audio router server is copied to the speech recognition server, which performs speech recognition on each sent message to detect the predetermined end-call phrase. Note that more than one end-call phrase may be predetermined, and the speech recognition server may detect any of the predetermined phrases. If such a phrase is detected, the speech recognition server signals the audio router server, which closes down the call. At the same time, the speech recognition server may signal the recipient determination server 16, which may sends "call-ended" signal to the visual selector client 104 at the user terminal. In such a case the visual selector client 104 then displays a "call ended" message to the user on the display screen of the mobile station 10. An audio "call ended" output using a synthesised or stored audio message may similarly be sent to the mobile terminal.

In the above-described embodiment, mobiles A and B communicate with the servers using a cellular wireless network. In alternative embodiments, a non-cellular wireless access network such as wireless LAN, Wi-Fi and Bluetooth could be used instead. In further alternative embodiments, one or both terminals could be fixed (e.g. a personal computer).

Unless the context clearly requires otherwise, throughout the description and the claims, the words "comprise", "comprising" and the like are to be construed in an inclusive as opposed to an exclusive or exhaustive sense; that is to say, in the sense of "including, but not limited to".